How Much of Native Language Identification Is Corpus Artifact? A Cross-Corpus Audit of SVMs and LLMs on TOEFL11 and ICLEv3

Shima Rahimi
Independent Researcher, Iran
shimarahimi1988@gmail.com

Abstract

We ask how much of the accuracy in native language identification (NLI) comes from real language signals and how much from problems in the corpus. We compare a supervised SVM, trained only on TOEFL11, with a zero-shot LLM (gpt-4o-mini) on eight native languages, testing both on TOEFL11 and on the separately collected ICLEv3. On TOEFL11 the LLM gains 35.6 to 48.5 percentage points from full text over content-masked text, depending on prompt format and filter tier, while the SVM gains 13.6 to 16.0. Normalising for chance, the SVM keeps about 77% of its above-chance accuracy under masking and the LLM about 4%, which matches the LLaMA-3.1-8B result of Uluslu and Schneider (2025) rather than their 70B. Letting the LLM reason first lowers its full-text accuracy but does not raise its masked accuracy. On ICLE, corpus-ID headers and essays that mention the writer's own country or language both inflate the LLM's accuracy, while topics tied to one language make essays harder for it. The SVM is far less affected, and its accuracy is unchanged by the header. After both filters, 71 essays still reach 91.5% (chance: 12.5%), an upper bound because the topic filter removes the essays the LLM finds harder.

1 Introduction

Native Language Identification (NLI) is the task of inferring a writer's first language (L1) from their writing in a second language (L2). The idea behind this task is cross-linguistic influence: a writer's first language affects how they write in the second language and leaves some indication of the L1 in their texts (Yu and Odlin, 2016). One main application of NLI is computer-assisted language learning software. If a system can identify the native language of a learner, it can predict the errors the learner might make in their writing, help in providing special materials for that specific learner so that the learner can work on those errors, and explain the differences between L1 and L2. Similar tools are already being developed using learner corpora, and L1-aware feedback is one of the practical promises of ICALL. Another application of NLI is forensic linguistics, in which an inference about a writer's background can help narrow a set of possible authors.

This application is the main reason why it matters whether an NLI system works for the right reason. Teachers can use the cross-linguistic patterns that a valid and accurate NLI system provides in identifying a learner's L1. A system that uses shortcuts, like a writer mentioning their own country or language, or a whole class being assigned to write about the same novel, looks the same as a system that uses cross-linguistic patterns if we consider only accuracy, but the first one does not explain anything about language, and its feedback does not convey useful information for a new learner or a new classroom. This paper is about the difference between those systems.

Two problems in current NLI research still need attention. First, most NLI studies test their models using cross-validation within a single corpus. This can cause the model to confuse the L1 signals with corpus-specific patterns that might correlate with L1 in that specific dataset. Brooke and Hirst (2013) showed that topic bias alone in the ICLE corpus can make NLI performance look much stronger than it really is. Second, large language models (LLMs) are now widely used, including for NLI (Zhang and Salle, 2023). The problem is that they have access to vast training data and general world knowledge, so they might guess a writer's L1 from hints in the text, such as a country name, that a model trained only on the target corpus could never use. Uluslu and Schneider (2025) studied the reliance of the model on lexical shortcuts directly, by masking the

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 1–15, November 19–20, 2026, University of Tartu Library

content words with part-of-speech tags and comparing LLaMA-3.1 on full text against masked text. The accuracy decreased greatly. Their study covered four closely related European languages, and we describe it in more detail in Section 2.

This paper asks two questions. RQ1: in a real cross-corpus setting, training only on TOEFL11 and testing on both TOEFL11 and the separately collected ICLEv3 corpus, across eight typologically different native languages, how much do a traditional SVM and a zero-shot LLM (gpt-4o-mini) rely on structure versus lexical content? Does letting the LLM reason freely, instead of forcing a fixed output format, change this reliance? RQ2: how much of the LLM's apparent accuracy on full-text ICLE comes from the real L1 signal, and how much comes from specific and measurable corpus problems, namely corpus-ID headers that state each essay's country of origin, essays mentioning the writer's own country or language, and essay topics that correlate with L1 because of how ICLE collects essays by institution? Do the SVM and the LLM differ in how much they are affected by each problem, and if so, why?

This paper makes three contributions. First, a cross-corpus comparison of an SVM and an LLM across eight languages from different language families, more than the four related languages Uluslu and Schneider (2025) used. Second, a comparison between a forced-format prompt and a free-reasoning prompt for the LLM, to test whether the format itself changes how much the model relies on lexical shortcuts. Third, three measured audits of corpus bias in ICLEv3: one on a corpus-ID header that states each essay's country of origin, one on essays mentioning the writer's own country or language, and one that tests whether Brooke and Hirst's (2013) topic-bias finding transfers from an SVM to a zero-shot LLM, and finds that in our data it does not. We show that the SVM and the LLM are affected differently by these problems, and based on how each system is built we suggest a reason for this difference.

2 Related Work

Classical NLI. Koppel et al. (2005) were among the first to treat NLI as an NLP task. They trained SVM models using part-of-speech bigrams, character n-grams, function words, and spelling-error types, and identified the L1 of writers from five language backgrounds with 80% accuracy, using the first version of ICLE. The field later built shared tasks around this problem. Tetreault et al. (2013) ran the first NLI shared task at the BEA-8 workshop and introduced the TOEFL11 corpus (Blanchard et al., 2013), which is the corpus we train on here and describe in Section 3. Malmasi et al. (2017) ran the second shared task, which combined written and spoken data, and drew submissions that used many feature types together, such as spelling errors, word and lemma n-grams, character n-grams, dependency trees, and morphological features.

Cross-corpus NLI and corpus bias. Malmasi and Dras (2015) showed that NLI systems trained and tested within one corpus through cross-validation do not always perform well on a second, separately collected corpus. This makes cross-corpus testing a harder and more informative test than cross-validation alone. Brooke and Hirst (2013) explained one reason why. Using four languages from ICLE (French, Spanish, Chinese, Japanese), they compared accuracy under a random train/test split against accuracy under a topic-controlled split, where no topic appears in both the training and test sets. Every feature type they tested dropped in accuracy under the topic-controlled split. This included part-of-speech n-grams (92.1% to 85.0%) and function words (85.2% to 77.1%), which earlier work had assumed were safe from topic bias. Character n-grams dropped the most (82.9% to 50.2%), and even the combined feature set dropped from 96.5% to 86.1% (all differences significant at $p < .001$). Brooke and Hirst argue this problem is deeper than shared topic words. In ICLE, each institution assigned its own essay topics, so topic is linked to both register and native language. Jarvis and Paquot (2012) suggested a way to fix this: remove four language groups (Chinese, Japanese, Turkish, and Tswana) that show the most variation in topic and skill level, and remove n-grams that also appear in the essay prompts. Brooke and Hirst point out that these four groups are exactly the non-European languages in ICLE, so removing them would limit any resulting system to European languages only. This is a problem for a task that should work across many languages. In this paper, we follow Brooke and Hirst's view rather than Jarvis and Paquot's filtering method. We keep eight typologically different languages, including Japanese, Korean, and Turkish, and instead offer an audit method (Section 6) as another way to address the same concern, without removing

non-European languages.

LLM-era NLI. Zhang and Salle (2023) showed that general-purpose LLMs can reach strong NLI accuracy in a zero-shot setting, without any task-specific training, while also giving natural-language explanations of their reasoning. Ng and Markov (2025) compared open-source and closed-source LLMs on NLI and found that open-source models do not match the accuracy of closed-source models without fine-tuning, but reach comparable accuracy once fine-tuned on labelled data. Most relevant to our work, Uluslu and Schneider (2025) tested whether an LLM's strong NLI performance reflects real structural analysis or reliance on content. They replaced content words with part-of-speech tags, keeping function words and punctuation, and compared LLaMA-3.1-8B and -70B on masked text against full text. On their TOEFL11 subset (TOEFL4: French, German, Italian, Spanish), the 70B model reached 96.5% accuracy zero-shot on full text, which dropped to 55.2% under masking. This is well above the 25% chance rate for their four-way task, but still a large drop. After fine-tuning an 8B model on the masked condition, performance recovered to 89.2% in-domain and 90.3% under cross-domain transfer to their ICLE subset, which shows the structural signal is considerable once the model is trained for it. The authors note that their own study is limited to four closely related European languages.

Gap. To our knowledge, no earlier work has applied this kind of masking-based audit to more than four typologically related languages in a true cross-corpus setting, or tested directly whether Brooke and Hirst's long-standing, SVM-based finding about ICLE's topic bias still holds for a zero-shot LLM that was never trained on ICLE. Goswami et al. (2024), in their survey of the NLI field, treat topic bias as a significant confounder for word-based features and note that NLI models may not generalise across datasets. This paper addresses cross-corpus generalisation and shortcut auditing together, using one experimental design. The two NLI shared tasks (Tetreault et al., 2013; Malmasi et al., 2017) are a useful point of comparison. Both were run as within-corpus tasks with cross-validation, which is the same setting that Malmasi and Dras (2015) later showed does not carry over to a second corpus collected separately. Neither shared task tested its systems with a masking-based check for lexical shortcuts, because this concern only became important once general-purpose LLMs, which have broad world knowledge, started being used for the task (Zhang and Salle, 2023). The gap we address is not only about using more languages than Uluslu and Schneider. It is also about running a shortcut check that was not a common concern when the main way to test NLI systems was cross-validation within TOEFL11 or ICLE.

3 Data

TOEFL11 (Blanchard et al., 2013) has 1,100 essays for each of eleven native languages. All essays respond to the same set of eight prompts, under standard exam-timing conditions. This paper uses eight of the eleven language groups: German, French, Italian, Japanese, Korean, Spanish, Turkish, and Chinese, split into 1,000 training and 100 test essays per language (8,000 training, 800 test).

ICLEv3 (Granger et al., 2020) has argumentative essays from more advanced learners across 25 native-language backgrounds. Unlike TOEFL11, ICLEv3 essays do not share one fixed set of prompts; topics vary by the contributing institution. This is an important design difference, and it motivates the audit in Section 6. This paper uses the same eight language groups from ICLEv3 as an external test set. The counts per language are: German 445, French 314, Italian 398, Japanese 366, Korean 400, Spanish 250, Turkish 286, Chinese 160 (2,619 essays in total).

Because TOEFL11 uses a shared, timed set of prompts and ICLEv3 does not, the two corpora are not equally exposed to topic-based shortcuts or to country and language names in the essay text. We state this design difference here, and we measure its effect directly in Section 6.

4 Methods

4.1 Part-of-speech masking

Following the general idea in Uluslu and Schneider (2025), though not their exact masking setup, we replace content words with part-of-speech tags using spaCy's `en_core_web_sm` model. NOUN, PROPN, VERB, ADJ, and ADV are mapped to single letters (N, P, V, A, R). All other tokens, including function words and punctuation, are left in place, so the masked version keeps the word order and the function words of the essay while removing

its content words. The full-text version keeps paragraph and sentence boundary markers (<P>, <S>). In the masked version spaCy splits each marker into three tokens, so <P> comes out as < P > or a similar form. The boundaries stay visible, but not always which kind of boundary it was. We found this after the runs were complete; a custom tokenizer rule would fix it. Both systems received the same masked text, but the masked LLM prompt describes the markers as intact and also uses P for proper nouns, so the damage may have affected the LLM more than the SVM. We also checked that masking keeps token counts aligned: for all 11,419 essays the masked text has the same number of tokens as the original.

4.2 SVM

We use TF-IDF features (word 1–3 grams, character 2–5 grams, with 5,000 and 10,000 maximum features) with a LinearSVC classifier. Both the masked and full-text vectorizers are fit only on the 8,000 TOEFL11 training essays, never on the TOEFL11 test essays and never on ICLE, in either condition. The classifier is likewise fit only on those 8,000 training essays. Every accuracy we report for the SVM is therefore measured on essays the model did not see during training. This means the comparison between masked and full-text isolates the effect of masking alone. It also matters for a finding in Section 6: since ICLE text is only ever transformed, not used for training, the SVM's vocabulary cannot pick up patterns that appear only in ICLE's raw text.

4.3 LLM

We use gpt-4o-mini in a zero-shot setting, under two prompt formats: a forced-JSON format with a small, 80-token output limit, and a free-text reasoning format, where the model reasons first and then gives a final answer on a separate line, with a 450-token limit. Both formats are tested on the same documents, which lets us test directly whether asking the model to reason changes how much it relies on structure versus content.

We called the model through the OpenAI API using the model name `gpt-4o-mini`. This is an alias rather than a dated snapshot, so the weights it serves can change over time. All four evaluation runs were made on 17 July 2026, and every API response in our logs records the model name `gpt-4o-mini`. Temperature was set to 0 in all runs.

4.4 Evaluation set

All systems and conditions are tested on the same 800-document set: 50 essays per language label per partition (`iclev3_external_test`, `toefl11_test`), reused across every SVM and LLM condition. Using the same set for every condition makes the comparisons in Section 5 fair.

The TOEFL11 half of this set is drawn only from the 800-essay TOEFL11 test split, never from the 8,000 essays used to train the SVM. We checked this directly: no document appears in both the training split and the evaluation set. The TOEFL11 numbers we report are therefore held-out in-domain results, and the ICLEv3 numbers are fully external, since no ICLE essay is used at any point in training.

The 50 essays per label were drawn at random from each test set with a fixed seed. We report results at three filter tiers. The *raw* tier uses all essays, with the ICLE corpus-ID header removed (Section 6.1). The *clean* tier also removes essays that mention their own L1's country or language (Section 6.2). The *doubly-clean* tier also removes ICLE essays on language-exclusive topics (Section 6.3). The topic filter uses only ICLE metadata, so for TOEFL11 the clean and doubly-clean tiers are identical.

4.5 Reproducibility

All code used in this paper is publicly available: the preprocessing and masking pipeline, the SVM and LLM evaluation, the three corpus audits in Section 6, and the scripts for the significance tests reported here. The code is available at `https://github.com/ShmRhm/nli-corpus-artifact-audit`.

We cannot release the corpora themselves. TOEFL11 and ICLEv3 are both licensed, so the released code assumes the reader has obtained them independently and documents the file layout it expects. There are two further limits on exact reproduction. The LLM results depend on the model version served by the OpenAI API at the time of our runs, and we cannot guarantee that the same model will return the same predictions later. The SVM results are deterministic given the same input corpus and a fixed random seed. The audits in Section 6 are deterministic and depend only on the corpus and the corpus metadata, so they can be re-run and checked directly by anyone with a licensed copy of ICLEv3.

5 Results

5.1 Main results

Table 1 shows the main results on the balanced 800-document evaluation set (50 essays per language per partition), which every system and condition shares. Two further SVM numbers are not in Table 1: on the complete, unbalanced sets, the SVM's full-text model reached 74.6% accuracy across all 2,619 ICLE essays and 83.0% across all 800 TOEFL11 test essays. We give these separately because they come from a larger set than the balanced 400-per-partition comparison the LLM is run on, so they are not directly comparable to the numbers in the table.

These results match the pattern discussed in Section 6. On the balanced set in Table 1, the LLM's full-text accuracy is higher on ICLE than on TOEFL11 (87.0% vs. 63.3%), while the SVM shows the opposite pattern (76.5% vs. 83.5%, lower on ICLE). This is an early sign that something drives the LLM's high ICLE accuracy that is not equally available to the SVM.

5.2 Reasoning format

Masked-text accuracy is 14.9% under forced-JSON and 14.6% under the reasoning format, a small decrease rather than an increase. The same small, flat, or slightly negative pattern appears on both partitions (ICLE: 15.0% to 14.8%; TOEFL11: 14.8% to 14.5%). Both formats stay close to the 12.5% chance rate for eight-way classification. Letting the LLM reason before answering does not recover masked-text performance under either format we tested. Uluslu et al. (2024) report a similar pattern on a different dataset: redacting named entities lowered an LLM's NLI F1 from 0.69 to 0.48 under a plain classification prompt, while a chain-of-thought prompt scored lower on unredacted text (0.54) but lost much less under redaction (0.47). Our reasoning format behaves in a similar way: it lowers full-text accuracy (Table 1) without raising masked accuracy. The reasoning prompt also adds a forensic-linguist role and a list of clue types (Appendix A), so this compares two prompts rather than two output formats alone. The near-chance masked result is not caused by the strict output format, because it remains when the model is free to reason. However, it does not tell us whether the model has no structural signal to work with, or is simply weak on this kind of input (Section 8).

5.3 Shortcut gaps on TOEFL11

Our main comparison is the gap between full-text and masked-text accuracy on the TOEFL11 partition, not on ICLE, and not on the unfiltered ICLE score. TOEFL11's shared-prompt, timed design controls for the topic and self-disclosure problems studied in Section 6, while ICLE's design does not.

Table 2 shows these gaps. We prefer the clean tier over the raw tier as our main number. The self-disclosure filter (Section 6.2) removes about 11 to 12 percent of the TOEFL11 test set as well, not only ICLE essays. This means raw TOEFL11 accuracy is already inflated by the same problem the audit is meant to catch. The clean and doubly-clean tiers are identical for TOEFL11, because the topic-exclusivity filter (Section 6.3) is based only on ICLE metadata and does not remove any TOEFL11 rows. The gap is stable from the clean tier onward, but not across every filter, so the filter matters even for TOEFL11, the corpus we use as a control.

All three raw-tier gaps are statistically significant. Because full-text and masked-text accuracy are measured on the same 400 TOEFL11 documents for each system and format, we test this with a paired exact McNemar test on the discordant pairs: LLM forced-JSON (197 documents correct only in the full-text condition versus 3 correct only in the masked condition, $p = 1.7 \times 10^{-54}$), LLM reasoning (177 versus 9, $p = 1.3 \times 10^{-41}$), and SVM (87 versus 23, $p = 6.0 \times 10^{-10}$). All three results are significant well beyond usual thresholds. The SVM's smaller effect size shows up as a less extreme p-value than the LLM's, though still highly significant.

The clean-tier gaps are also significant, using the same paired McNemar test restricted to the 354 TOEFL11 documents that survive the self-disclosure filter (Section 6.2): LLM forced-JSON (155 full-text-only versus 3 masked-only, $p = 3.6 \times 10^{-42}$), LLM reasoning (135 versus 9, $p = 5.5 \times 10^{-30}$), and SVM (70 versus 22, $p = 5.3 \times 10^{-7}$). Removing the self-disclosure confound shrinks the discordant-pair counts, as expected, but does not remove the effect: all three gaps remain significant well beyond conventional thresholds.

5.4 What the errors look like

Accuracy alone does not show whether a system fails in a linguistically sensible way. We therefore looked at which languages each system confuses on

System	Format	Condition	n (all)	Acc. (all)	n (ICLE)	Acc. (ICLE)	n (TOEFL11)	Acc. (TOEFL11)
SVM	—	full text	800	80.0%	400	76.5%	400	83.5%
SVM	—	masked	800	63.4%	400	59.3%	400	67.5%
LLM	forced-JSON	full text	800	75.1%	400	87.0%	400	63.3%
LLM	forced-JSON	masked	800	14.9%	400	15.0%	400	14.8%
LLM	reasoning	full text	798*	67.7%	398*	78.9%	400	56.5%
LLM	reasoning	masked	800	14.6%	400	14.8%	400	14.5%

Table 1: Main results across systems, prompt formats, and masking conditions on the 800-document evaluation set. *Two ICLE documents were dropped from the full-text reasoning condition because of a parse or API error, and were logged as such at evaluation time.

System	Format	Tier	Full	Masked	Gap
LLM	f-JSON	raw	63.3%	14.8%	**48.5**
LLM	f-JSON	clean	58.5%	15.5%	**42.9**
LLM	reasoning	raw	56.5%	14.5%	**42.0**
LLM	reasoning	clean	51.1%	15.5%	**35.6**
SVM	—	raw	83.5%	67.5%	**16.0**
SVM	—	clean	81.9%	68.4%	**13.6**

Table 2: Full-text vs. masked-text accuracy gap (in percentage points) on the TOEFL11 partition. "f-JSON" = forced-JSON. "clean" = "doubly-clean" tier (see Section 6).

the TOEFL11 partition, where prompts are shared and topic bias is controlled.

The SVM's errors follow typological lines. In the masked condition, where only structure is left, its most frequent confusions are Korean predicted as Japanese and Japanese predicted as Korean (9 essays each), together 18 of the 130 errors it makes on this partition, or about 14%. Next comes French predicted as German (6), then five pairs at 5 essays each: Japanese as Turkish, Japanese as Chinese, French as Spanish, Turkish as French and Turkish as Spanish. Japanese, Korean and Turkish are all verb-final and agglutinating, and French, German and Spanish are all Indo-European, so most of these errors stay within groups of related or structurally similar languages. Three do not. Japanese and Chinese differ in word order and morphology, although neither language has articles, and article use survives masking. Turkish predicted as French or Spanish does not fit the pattern at all. The same broad picture holds in the full-text condition, where the largest confusions on this partition are French as German (6), Korean as Japanese (5), Japanese as Korean (4) and Italian as German (4), out of 66 errors. An NLI system that mixes up Korean with Japanese, and French with Spanish, is making the kind of mistake that a linguist would expect from a system reading grammatical structure.

The LLM under masking fails in a different way. It does not spread its errors across typologically close neighbours; it stops using most of the label set. On masked TOEFL11 text under the forced-JSON format, it predicts Spanish for 298 of 400 essays (74.5%) and uses only four of the eight labels at all. Under the reasoning format it predicts Turkish for 190 of 400 essays (47.5%) and uses five labels. For comparison, the same model on full-text TOEFL11 uses all eight labels, with its most frequent prediction accounting for 41% of essays. The near-chance masked accuracy in Table 1 does not come from the model making fine-grained mistakes between similar languages. It comes from the model largely abandoning the task and falling back on a default answer.

Both systems lose accuracy when content words are removed, but only the SVM keeps making linguistically structured errors. The LLM's behaviour under masking looks less like weak structural analysis and more like the model abandoning the task. Section 8 discusses what this means for the model's structural ability.

6 Corpus-Bias Audit

6.1 Document-ID headers

The raw ICLEv3 text files begin with a corpus-ID header, for example `<ICLE-FR-UCL-0027.2>`. This header is part of the corpus file, not of the essay, but it encodes the country code and the institution code of each essay, so it states the answer to the classification task in the first few characters of the input. It is present in all 2,619 ICLEv3 essays, and TOEFL11 files never contain it.

In our first runs the header was passed to both systems as part of the essay text. This was a preprocessing oversight, but an easy one to make, since the header sits in the same file as the essay with no separator and nothing in the corpus documentation warns that it encodes the label. We report its effect because it shows how readily an LLM uses leaked metadata, and so that other users of ICLEv3 can

check for it.

We remove the header from the ICLE text and re-run both systems. This gives a clean comparison, because only one thing changes between the two runs. The effect on the two systems is very different. The LLM's full-text ICLE accuracy drops from 98.8% to 87.0% under the forced-JSON format, and from 98.0% to 78.9% under the reasoning format. The SVM's full-text ICLE accuracy does not change at all: it is 76.5% both before and after, on the balanced evaluation set.

We checked the SVM's fitted TF-IDF vocabularies directly to confirm why. Both vectorizers are fit only on TOEFL11 training text and only ever transform ICLE text (Section 4.2). To confirm this, we passed the header string alone through the fitted vectorizers: it produces no word features at all, and six character n-grams (`ic`, `cl`, `cle`, `le`, `fr`, `uc`). The header therefore carries almost no country evidence the SVM can act on: even `fr`, the country code in this example, is common in ordinary words such as *from* and *free*. This is why the SVM's accuracy is unchanged whether or not the header is present. The LLM reads the raw text at test time and does not make this distinction, so it can use the header's country code wherever it appears.

This is the clearest single case in our data of the wider pattern we report in this paper: the same corpus artifact costs one system between 11.8 and 19.1 accuracy points, depending on the prompt format, and the other nothing at all, for a reason that follows from how each system is built rather than from what each system knows about language.

6.2 Self-disclosure

We use a keyword search to flag essays that mention their own L1's country or language anywhere in the text. As the hand-check later in this section shows, some of these essays state the writer's nationality directly while others only discuss the country. The rate of flagging is uneven across languages, and much higher in ICLE than in TOEFL11 (Table 3).

Label	ICLE (external_test)	TOEFL11 (test)
JPN	71.3%	24.0%
KOR	55.5%	21.0%
SPA	25.2%	4.0%
ITA	25.1%	19.0%
TUR	24.1%	5.0%
ZHO	23.8%	5.0%
FRA	21.0%	11.0%
DEU	18.2%	8.0%

Table 3: Self-disclosure rate by native-language label.

On the 800-document evaluation set, the LLM's full-text forced-JSON accuracy is 67.9% on essays without a self-disclosure flag, and 99.5% on essays with one ($n = 616$ vs. 184). Across all 3,419 essays the SVM was run on (the 2,619 ICLEv3 external test essays plus the 800 TOEFL11 test essays), its full-text model shows a smaller but still clear gap: 73.4% without disclosure versus 84.5% with it ($n = 2{,}422$ vs. 997). Both splits are statistically significant: a chi-square test of self-disclosure flag against correctness gives $\chi^2(1) = 74.0$, $p = 7.7 \times 10^{-18}$ for the LLM and $\chi^2(1) = 47.8$, $p = 4.7 \times 10^{-12}$ for the SVM. The LLM's accuracy gap (31.6 points) is about three times the SVM's (11.1 points); the chi-square values themselves are not comparable, since the two sets differ in size. This is consistent with the pattern in Section 5: the LLM is more sensitive to this particular shortcut than the SVM is, though both are affected well beyond chance.

We checked this filter by hand on a random sample of 48 flagged essays drawn from both corpora. In 15 of them (31%) the writer links themselves to the country or language directly, as in *"My country, Korea"* or *"living in Italy, I know…"*. In 32 the essay discusses its own country or language without the writer saying they belong to it, as in *"In Japan, the death penalty is decided to be the ultimate punishment"*. The remaining one is a plain error: the phrase *"french fries"* in an essay by a French speaker, where the word has nothing to do with nationality. Uluslu et al. (2025) show the same effect in a controlled test: inserting a stereotype sentence about Italian coffee habits into an essay by a Spanish writer led an LLM to predict Italian, even though the prompt instructed it to disregard cultural references. This suggests that the effect comes from country- and culture-linked words in general, not from our keyword list. Our filter is better described as finding essays that mention their own L1's country or language than as finding direct self-disclosure. This does not change the accuracy differences above, which follow from the lexical cue being present in the text at all, but it does mean the filter is broader than the term "self-disclosure" suggests, and we use the term in this wider sense throughout.

The LLM is correct on 183 of the 184 flagged essays in the evaluation set, yet in only about a third of flagged essays does the writer say anything about

belonging to that country. The model is almost always right when an essay names its own L1's country or language, in any sentence and about anyone, whether or not the writer says they come from there. For a teaching tool this makes the problem harder: a learner does not have to write about their own background for the cue to appear.

The filter has two limits. It searches for the terms tied to each essay's own label, so it can only find cues that point the model toward the correct answer. A Spanish writer's essay about Japan is not flagged, even though the word *Japan* could push the model the wrong way. Our clean tier therefore still contains country and language names, just not the essay's own ones, and the 67.9% we report for unflagged essays may be lower than what the model would reach on text with no country names in it at all.

6.3 Topic exclusivity

ICLEv3's metadata includes a Title field, which records the essay prompt or title assigned to each text. We use each distinct title string directly as a topic identifier: we do no clustering or topic modelling, so two essays share a topic only if their Title fields match exactly. For every title we then compute the distribution of native-language labels across the essays carrying it, and flag an essay as topic-exclusive when a single label accounts for at least 80% of the essays under its title. Label shares are computed over all 2,619 ICLE essays, not over the filtered subset. Note that a title carried by a single essay satisfies the rule automatically. Of the 613 titles that meet it, 425 are used by only one essay, and those account for 425 of the 2,058 essays flagged, so "topic-exclusive" partly means "has a title no other essay shares". This is a deliberately simple rule, and Section 8 discusses what it misses. We compute this on the self-disclosure-clean ICLE subset (262 essays: 71 non-exclusive, 191 exclusive). The LLM's full-text forced-JSON accuracy is 91.5% on the 71 non-exclusive essays and 76.4% on the 191 exclusive essays. The exclusive-topic essays are actually harder for the LLM, not easier. This split is statistically significant but visibly weaker than the self-disclosure split above: $\chi^2(1) = 6.6$, $p = 0.010$ (Fisher's exact test, which is a better fit for the small cell counts here, gives $p = 0.0050$, so the effect is not just a side effect of the chi-square approximation). We treat this as an open question. Topic exclusivity is strongly linked to language label itself (German, Japanese, and Korean essays are 97 to 100 percent topic-exclusive, while French and Spanish are 32 to 44 percent), so any claim about "topic" driving this pattern needs to account for label distribution first.

We therefore checked whether label distribution alone explains the gap. It does not. In this subset, no German, Japanese, or Korean essay falls in the non-exclusive group, and those are among the labels the LLM handles worst (Korean 52.6%, Japanese 63.6%), so their absence should favour the non-exclusive group. But when we restrict the comparison to the five labels that appear in both groups (French, Italian, Spanish, Turkish, Chinese), the gap remains almost unchanged: 91.5% on non-exclusive essays ($n = 71$) against 76.4% on exclusive essays ($n = 123$). Whatever makes topic-exclusive essays harder for the LLM is therefore not only a matter of which languages they belong to. We do not have an explanation for it.

We ran the same split for the SVM. Here we use the full self-disclosure-clean ICLE set ($n = 1{,}719$), which is larger than the LLM's 262 essays because the SVM was scored on all of ICLE. Masked accuracy is 56.6% on non-exclusive essays ($n = 429$) against 58.9% on exclusive essays ($n = 1{,}290$). Full-text accuracy is 71.8% against 69.3%. Both gaps are small, and neither is significant (masked: $\chi^2(1) = 0.59$, $p = 0.44$; full text: $\chi^2(1) = 0.84$, $p = 0.36$). The SVM is not sensitive to this split. For the LLM, the topic effect is significant but much weaker than the self-disclosure effect.

6.4 Remaining signal

We removed the essays flagged for mentioning their own country or language, and also the essays on language-exclusive topics. The 71 ICLE essays left show 91.5% accuracy for the LLM's full-text forced-JSON condition, well above the 12.5% chance rate for eight-way classification, though $n = 71$ is a small sample (approximate 95% binomial confidence interval: 83 to 96 percent). This number needs care, because these 71 essays are not an average sample of the clean ICLE set. They are the same non-exclusive group reported in Section 6.3, where the LLM reached 91.5% against 76.4% on the 191 exclusive-topic essays. Because the topic filter takes out the essays the model found harder, 91.5% is the high end of the clean set, not an average of it, and should be read as an upper bound. A signal that stays well above chance af-

ter the self-disclosure and topic filters is what we would expect if some real cross-linguistic signal is present in the text, which is the basic assumption behind NLI. We do not read this residual as a sign that the audits failed, but as a limit on how much of the LLM's ICLE accuracy these audits can account for. Our audits explain part of the LLM's accuracy, but not all of it. We ran a further check. This time we did not look for essays that name their institution, but for essays whose words are much more common at one institution than in the rest of the corpus. For each institution with enough essays to test, we find short word tokens that appear far more often in that institution's essays than elsewhere, then flag any essay that contains at least one such token. In practice this mostly finds shared reading vocabulary, for example character and author names from a novel that a whole class wrote about, rather than any direct mention of the institution's name. We call this the institution-vocabulary check, not "institution self-mention," so as not to suggest the essays name their institution. This gives 74.7% accuracy without a flagged token ($n = 79$) and 83.1% with one ($n = 183$), computed on the same 262-essay self-disclosure-clean set as Section 6.3. Unlike the other splits we report, this difference is not statistically significant (chi-square test of institution-vocabulary flag against correctness: $\chi^2(1) = 2.0, p = 0.16$), so we treat it as suggestive only.

The two topic-based checks also point in opposite directions. Essays with institution-specific vocabulary are answered slightly better (83.1% against 74.7%), while essays on language-exclusive topics are answered worse (76.4% against 91.5%, Section 6.3). Only the second difference is significant, so we do not read this as two competing effects.

We also measured how closely this institution-vocabulary check is related to topic exclusivity, using the full 2,619-essay ICLE set. This test is between the two checks themselves, not against correctness, and it gives $\chi^2(1) = 159.0$, $p = 1.9 \times 10^{-36}$, $\phi = 0.246$. This shows a modest link between the two checks, so we do not treat the institution-vocabulary check as a fully separate third problem. Because its own accuracy split is not significant, we also do not treat it as support for the remaining signal. The two checks are related, since institution-specific words are themselves a kind of topic signal, just found from the data for each institution instead of read from the Title field used in Section 6.3.

6.5 Concrete examples

The audits above report numbers over many essays. To make them easier to picture, we give one real example of each, with any identifying detail removed.

Masking. In a TOEFL11 essay (label: Japanese), the first two sentences in full text are: *"I disagree. I have three opinion to support this reason."* After masking, the same text becomes: *"I V . I V three N to V this N ."* Each content word is replaced one for one (*disagree*→V, *have*→V, *opinion*→N, *support*→V, *reason*→N), while the function words, punctuation, and sentence structure stay the same. This shows why accuracy well above chance on masked text would point to real structural signal, not to the model using content words as a shortcut.

Self-disclosure. One ICLEv3 essay (label: Spanish) is flagged by our keyword search only because it contains the country name *"Spain"* in the text, which is not related to the essay's topic. This is the kind of essay the filter in Section 6.2 is meant to catch: one word matching the essay's own label is enough for the flag, and, as that section shows, usually enough for the model too, whether or not the writer says they come from there.

Institution vocabulary. A group of essays from one Belgian institution are flagged because they share words such as *briscoe*, *ramsay*, *lighthouse*, and *woolf*, which are character and author names from Virginia Woolf's *To the Lighthouse*. No essay in this group names its institution. The institution can still be guessed because a whole class was given the same novel to write about, so the shared book acts as a marker for the institution even though it was never meant to be one.

7 Discussion

The TOEFL11 gaps (Section 5) show that, even on a corpus built to control for topic and register, the zero-shot LLM depends on content words much more than the supervised SVM does: a 35.6 to 48.5 percentage-point gain from full text over masked text for the LLM, depending on prompt format and filter tier, against 13.6 to 16.0 points for the SVM. The smallest gap for the LLM is more than double the largest gap for the SVM. Uluslu and Schneider (2025) report both outcomes

in the same table, so the comparison depends on which of their models is used. Their LLaMA-3.1-70B keeps 55.2% zero-shot accuracy on masked TOEFL4 text, well above the 25% chance rate for their four-way task, while their LLaMA-3.1-8B falls to 26.0%, which is chance. Our zero-shot gpt-4o-mini reaches 14.8% on an eight-way task, close to its own 12.5% chance rate, so our result matches their smaller model rather than their larger one.

Because the two studies use a different number of classes, raw accuracy and raw gaps are not comparable. A measure that is comparable is the share of above-chance accuracy that survives masking, $(\text{masked} - \text{chance})/(\text{full} - \text{chance})$. On this measure, computed on the unfiltered TOEFL11 essays under the forced-JSON prompt so that it is comparable with their setup, our gpt-4o-mini keeps about 4% and our SVM keeps about 77%, against about 42% for their 70B model and about 3% for their 8B. Retention differs sharply between two sizes of the same model family and is similar for the two smaller models, which points to capacity rather than to a general property of large language models. Since gpt-4o-mini's size is not public we can say our result is consistent with that reading, not explained by it. The comparison is still not between identical situations (different models, a different number of classes, and a masking setup that is similar but not identical), so we do not claim that all LLMs rely on shortcuts in the same way. Their 70B model kept some structural signal under masking, while their 8B model and our gpt-4o-mini did not. Model capacity is the most likely reason, but our larger set of languages and our different masking setup may also play a part. Running open models of several sizes through the same pipeline would answer this question.

The masking example in Section 6.5 helps explain why the reasoning format does not close this gap. Under masking, the LLM sees a string like *"I V . I V three N to V this N ."*, which is grammatical form only, with no content words to give a clue. Asking the model to reason before it answers can only help if there is enough structural signal in that part-of-speech sequence, and Section 5 shows there is very little of it for gpt-4o-mini. The problem is not that the prompt forces the model to answer without reasoning; it is the input the model has to work from. This explains why masked-text accuracy stays close to chance in both prompt formats, and it also explains the error pattern in Section 5.4: the model does not make careful, meaningful mistakes between similar languages, it mostly stops using the label set and returns one default answer.

The restriction of the SVM's features to its TOEFL11 training vocabulary explains the difference between the SVM and the LLM on the corpus-ID header (Section 6.1). A corpus-ID string such as `<ICLE-FR-UCL-0027.2>` does not occur in TOEFL11 text, so it matches no word feature and only a handful of character n-grams that are common in ordinary English words (Section 6.1), while the LLM reads all the text it is given at test time. Since shared reading vocabulary is often ICLE-specific in the same way, the same argument explains part, but not all, of the institution-vocabulary effect.

It does not explain the self-disclosure result. Country and language names are ordinary English words that occur in TOEFL11 essays too, so they do have feature columns, and the SVM can use them. That is why Section 6.2 finds the SVM significantly affected by self-disclosure ($\chi^2(1) = 47.8$, $p = 4.7 \times 10^{-12}$), even though the effect is smaller than the LLM's. This reasoning therefore explains the header result, part of the institution-vocabulary result, and none of the self-disclosure result. This does not mean the SVM is a better NLI system in general: it was trained on in-domain data, while the LLM was used zero-shot. It is a narrower point about which kind of mistake each system might make, and it suggests that training only on corpus A and testing on corpus B is not only good practice in general, but also happens to keep this kind of corpus leakage out.

One practical suggestion follows from Section 6: full-text accuracy on ICLE, or on any untimed, variable-prompt learner corpus, should not be reported without a similar check for self-disclosure and topic correlation. Corpora with shared prompts, like TOEFL11, are structurally safer for this purpose, though Section 5 shows that even TOEFL11 is not fully free of self-disclosure leakage and benefits from the same check.

What this means for ICALL. These results matter for building L1-aware feedback into language-learning tools. A common way to build such a tool is to take a learner corpus, train or prompt a system to predict L1, and then use the prediction to select feedback. In a real tool the learner's first language can usually be asked directly, so what NLI adds is knowing which features carry the L1 signal,

since those are what feedback can point to. That makes the source of a system's accuracy more important than its level. Our results suggest two considerations for that pipeline. First, a high reported accuracy on a learner corpus is not by itself evidence that the system has learned anything about the learner's native language. On ICLE, a large part of the LLM's accuracy came from a corpus-ID header, from essays mentioning the writer's own country or language, and from class reading lists, none of which will be present when a real learner types an essay into a tool. A system adjusted and validated on those cues will work accurately in development but fail in real use. Second, the masking result shows that when the lexical cues are hidden, the accuracy of the LLM we tested falls close to chance, which is consistent with it not reading the grammatical patterns that cross-linguistic influence actually produces. Feedback based on such a prediction would not be reliable, and would not be explainable to a teacher or a learner in terms of any real property of the writing. We would therefore suggest that developers of ICALL systems that use L1 prediction audit their training and evaluation corpora in the way we describe here, and report accuracy on content-masked text alongside full-text accuracy, since the gap between the two is a simple, low-cost indicator of how much a system depends on clues that will not transfer to the classroom or to real educational use.

Uluslu and Schneider's measured amount of shortcut reliance came from a four-language, closely related subset of one corpus family, without a separate test for the kind of topic- or self-disclosure-based problems this paper measures. Part of the effect they report may therefore reflect corpus-specific problems in their evaluation data. We did not test their data, so this is only a hypothesis.

8 Limitations

This study uses one closed-source LLM, gpt-4o-mini. Results may not transfer to other model families or sizes. The same care is needed for the main comparison in this paper. Our LLM reaches 63.3% on full-text TOEFL11, below the 83.5% our own SVM reaches on the same essays (Table 1). A model that trails a linear classifier in-domain, and that under masking falls back on a single label for most essays (Section 5.4), may be showing general weakness on unfamiliar input rather than a specific reliance on lexical content. Our results are consistent with both readings, and separating them would need a stronger LLM run through the same pipeline.

Our design also does not separate a second confound. The two systems differ in supervision as well as in architecture and representation: the SVM is fit on 8,000 labelled TOEFL11 essays, while the LLM receives no task-specific adaptation at all. A zero-shot model may simply never have been given the chance to learn that part-of-speech sequences carry L1 information, which would produce the same masked-text collapse we report without any claim about what LLMs attend to in general. Three experiments would separate these factors, and we did not run them: few-shot prompting, to test whether a small number of examples recovers masked-text accuracy; an open-weights model such as LLaMA-3.1, which would also make the LLM side of this comparison reproducible in a way an API alias cannot; and a fine-tuned encoder such as XLM-RoBERTa, which would hold supervision constant and let the effect of architecture be read on its own. Until those are run, our results should be read as a measurement of what this zero-shot system does on this data, not as a general property of large language models. The LLaMA-3.1-70B model of Uluslu and Schneider (2025) kept much more of its accuracy under masking than gpt-4o-mini, while their 8B model kept about as little (Section 7). This points to model capacity, which we could not test with one closed model. Our LLM is also served through an API alias rather than fixed weights (Section 4.3), so running the same comparison on open-weights models with a fixed seed would make the LLM side of it reproducible in a way ours cannot be. We report statistical significance for each comparison: an exact McNemar test for the paired full-text against masked-text gaps in Section 5, and chi-square tests for the confound splits, which compare independent groups of essays (Sections 6.2, 6.3 and 6.4). Section 6.4 reports two of these tests: the accuracy split for the institution-vocabulary check is not significant, while the association between that check and topic exclusivity is. Section 6.3 also reports Fisher's exact test, because the cell counts there are small. The 80% threshold used for topic exclusivity in Section 6.3 is a simple rule, not a precise measure.

Our masking is coarse. Each content word be-

comes one of five letters, so its inflection is lost with it. In the example in Section 6.5, *"I have three opinion"* becomes *"I V three N"*, which hides the missing plural after a number, a common error for learners whose first language does not mark plural on nouns. Masking keeps some kinds of transfer visible, such as article and preposition use, pronouns, word order and punctuation, but removes lexical transfer (cognates, false friends, L1-influenced word choice), spelling and inflection. The masking gap is therefore an upper bound on shortcut use rather than a direct measure of it. Finer tags, or masking in steps starting with named entities, would show how much each kind of information contributes; Uluslu et al. (2025) find that removing named entities alone does not stop an LLM relying on surface cues.

Our topic audit also uses the Title field from the ICLEv3 metadata rather than topics inferred from the essay text. This is deliberate: the Title field is external to the essay body, so it does not carry the L1-correlated content that the audit is meant to detect. A topic model fit on the essays themselves would risk defining topics partly from the same signal we are trying to control for. The cost of this choice is coverage, since two essays on much the same subject can carry different titles and will not be grouped together. Embedding-based topic models such as Top2Vec (Angelov, 2020) could be used to test how much of the residual signal in Section 6.4 is explained by topical similarity that the Title field misses. We leave this to future work, and note that such an analysis would need to be read with the circularity above in mind.

The 71-essay ICLE subset that survives both audits still shows accuracy well above chance, which means our current audit does not fully explain the LLM's ICLE performance. That subset is small, and the confidence interval around its accuracy is wide (Section 6.4), so it can support the claim that some signal remains but not a precise estimate of how much. Answering this would need a larger cross-corpus set audited in the same way, which we hope to build in future work. This study covers eight of the languages shared between TOEFL11 and ICLEv3, chosen because both corpora include them, not through random sampling of all available languages.

9 Conclusion

For RQ1, the supervised SVM and the zero-shot LLM differ in how much they rely on lexical content, even on the TOEFL11 partition, which is built to control for corpus design problems. The LLM gains 35.6 to 48.5 percentage points from full text over masked text, while the SVM gains only 13.6 to 16.0 points. Each of these gaps is statistically significant on its own (Section 5); we did not test whether the LLM's gap is significantly larger than the SVM's, so we report the difference as a consistent ordering rather than as a tested contrast. Letting the LLM reason before answering does not meaningfully reduce this reliance. Our masked-text collapse matches the LLaMA-3.1-8B result reported by Uluslu and Schneider (2025), while their 70B model stayed well above chance, which points to model capacity rather than to a property of large language models in general (Section 7).

For RQ2, a real part of the LLM's apparent full-text accuracy on ICLE, though not all of it, comes from two measurable corpus problems: document-ID headers and self-disclosure (Sections 6.1 and 6.2). The third problem we audited, topic exclusivity (Section 6.3), does not inflate accuracy in our data; essays on language-exclusive topics are harder for the LLM, and we leave that unexplained. The SVM is much less affected by these problems, and in the case of the document-ID header not affected at all, because of how it is built rather than what it knows about language: its features and weights come only from TOEFL11. Some signal remains (91.5% accuracy on the 71 essays that survive both audits); it is well above chance and is not fully explained by our audits. Because the topic filter removes the essays the LLM found harder, this figure should be read as an upper bound on that unexplained signal rather than an average of it (Section 6.4). A further check for institution-specific vocabulary does not reach significance, so it neither supports nor weakens this residual; we report it because it is only moderately linked to topic exclusivity ($\phi = 0.246$) and so is not measuring the same thing twice. Together, these results suggest that NLI studies on untimed, variable-prompt learner corpora need this kind of audit before full-text accuracy can be read as evidence of real language transfer, and that this need applies differently depending on the system's design.

Ethics Statement

This study uses only licensed, de-identified learner corpora (TOEFL11, ICLEv3), obtained under their respective agreements. No new data involving human subjects was collected. All LLM evaluation used gpt-4o-mini through the OpenAI API: 3,200 API calls for the runs reported in Table 1 (800 documents times two text conditions times two prompt formats), which used 2,338,928 prompt tokens and 306,570 completion tokens, taken from the logs recorded with each API response. At gpt-4o-mini list pricing ($0.15 per million input tokens, $0.60 per million output tokens), this comes to about $0.53. SVM training and testing ran locally and needed no paid computing.

Acknowledgments

This paper builds on and substantially extends the author's 2023 thesis, *In-Corpus and Cross-Corpus Analysis of Native Language Identification Using Machine Learning*, supervised by Walter Daelemans and Ehsan Lotfi (Department of Linguistics, University of Antwerp). We thank them for their supervision of that earlier work.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☒ Generation from prompt

Details: An AI assistant (Claude, Anthropic) drafted and revised prose throughout the paper, including the abstract; suggested analyses and wrote the significance-test scripts; produced a first pass of the annotation in Section 6.2, which the author read in full and confirmed; and shortened and expanded text to meet the page limits. The research question, data preparation, pipeline and evaluation runs are the author's own work, and the author has verified every number reported here against the result files. Separately, gpt-4o-mini is an object of study in this paper, not a writing aid.

References

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Julian Brooke and Graeme Hirst. 2013. Native language detection with "cheap" learner corpora. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*, pages 37–47. Presses universitaires de Louvain, Louvain-la-Neuve.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *International Corpus of Learner English. Version 3*. Presses universitaires de Louvain, Louvain-la-Neuve.

Scott Jarvis and Magali Paquot. 2012. Exploring the role of n-grams in L1 identification. In Scott Jarvis and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, pages 71–105. Multilingual Matters, Bristol, UK.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. Robust native language identification through agentic decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8398–8414.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88.

Ahmet Yavuz Uluslu, Gerold Schneider, and Can Yildizli. 2024. Native language identification improves authorship attribution. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 297–303.

Liming Yu and Terence Odlin. 2016. *New perspectives on transfer in second language learning*, volume 92. Multilingual Matters.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

A Prompts

All four prompts are reproduced here exactly as used in the runs of 17 July 2026. Temperature was 0 in every run. The user message in all four conditions was `Essay:\n\n{text}`.

A.1 Forced-JSON format

Output was constrained with the OpenAI structured-output API using a strict `json_schema` with two required properties, `predicted_l1` (enumerated over the eight language codes) and `confidence_score` (a number between 0 and 1), and no additional properties. The completion limit was 80 tokens. Only `predicted_l1` is used in this paper.

Full-text condition.

```
You are an expert computational
linguist evaluating Native
Language Identification from
English essays written by
non-native speakers.

The text may contain <P> and
<S> tokens marking paragraph
and sentence boundaries; treat
them as structural formatting
markers, not content.

Choose exactly one
native-language code from:
DEU, FRA, ITA, JPN, KOR, SPA,
TUR, ZHO.

Return only the requested JSON
object.
```

Masked condition.

```
You are an expert computational
linguist evaluating Native
Language Identification from
content-word-masked English
essays.

The essay uses these
placeholders:
- N = noun
- P = proper noun
- V = verb
- A = adjective
- R = adverb

Function words, punctuation,
pronouns, determiners,
auxiliaries, conjunctions,
prepositions, and structural
markers such as <P> and <S>
remain visible.

Choose exactly one
native-language code from:
DEU, FRA, ITA, JPN, KOR, SPA,
TUR, ZHO.

Return only the requested JSON
object.
```

A.2 Free-reasoning format

The completion limit was 450 tokens. The final label was read off with the regular expression `FINAL_ANSWER\s*:\s*(DEU|...|ZHO)`; a response with no match was logged as a parse error and excluded. The prompt was assembled from an introduction, a boundary-marker legend, a part-of-speech legend used only in the masked condition, and a fixed instruction block.

```
You are an expert computational
linguist and forensic linguist
performing Native Language
Identification (NLI) on English
essays written by non-native
speakers.

The text may contain <P> and
<S> tokens marking paragraph
and sentence boundaries; treat
them as structural formatting
markers, not content.

[masked condition only:]
The essay's content words
have been replaced with these
placeholders:
- N = noun
- P = proper noun
- V = verb
- A = adjective
- R = adverb
Function words, punctuation,
pronouns, determiners,
auxiliaries, conjunctions, and
prepositions remain in their
original form.

Think step by step.  Identify
concrete linguistic clues -
```

word order, article/preposition use, function-word patterns, syntactic structure, collocations, or spelling/grammar patterns - that point to a specific native language. Write out your reasoning in a short paragraph.

After your reasoning, end your response with exactly one line in this exact format, with no extra text after it:
FINAL_ANSWER: <CODE>

<CODE> must be exactly one of:
DEU, FRA, ITA, JPN, KOR, SPA, TUR, ZHO.